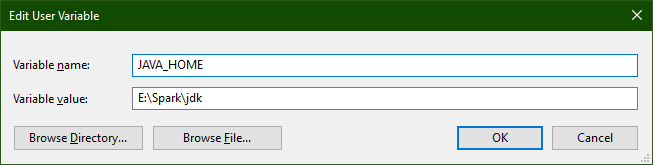
Setup Spark on Windows

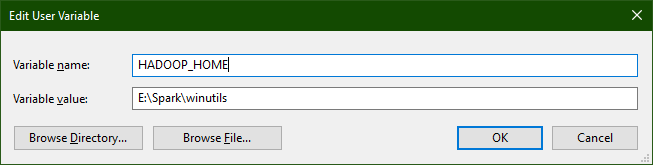
# Download **Jdk** version 8

Installed to folder E:\Spark\jdk and set user Environment variable as show below.

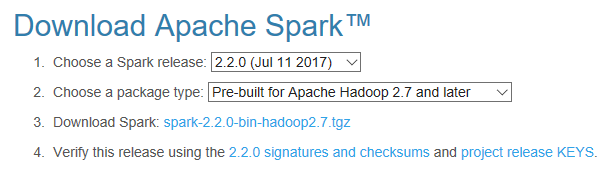


# Download **winutils.exe** file

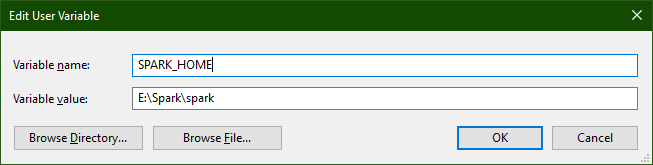
Copied to folder E:\Spark\winutils\bin and set user Environment variable as show below.



# Download **Spark** pre-built version



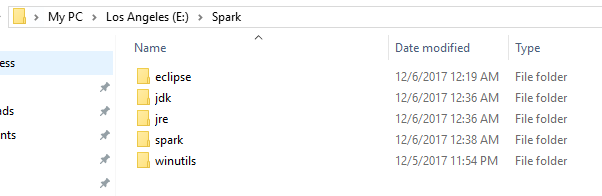
Unzipped to folder E:\Spark\spark and set user Environment variable as show below.



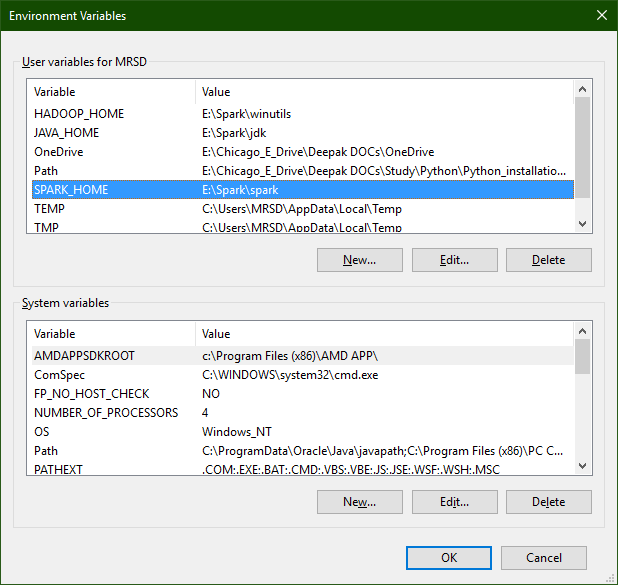
# Download **Scala** IDE

Extracted to folder E:\Spark\eclipse

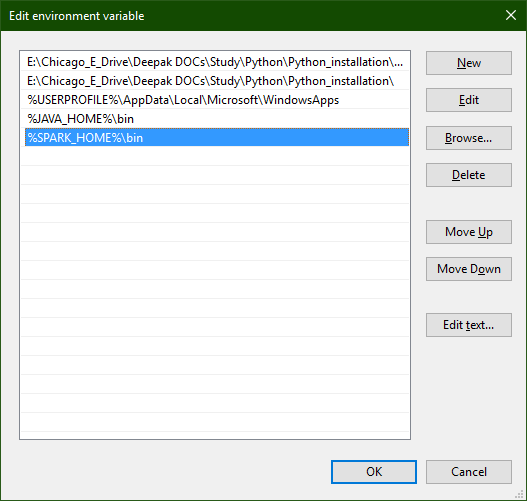
# Folder Structure



# User Environment Variables



# Path Variable



# Start-up spark-shell

Microsoft Windows [Version 10.0.15063]

(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\MRSD>cd e:

E:\

E:\>cd Spark\spark

E:\Spark\spark>spark-shell

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

17/12/06 00:38:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/12/06 00:38:33 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/E:/Spark/spark/bin/../jars/datanucleus-rdbms-3.2.9.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/E:/Spark/spark/jars/datanucleus-rdbms-3.2.9.jar."

17/12/06 00:38:33 WARN General: Plugin (Bundle) "org.datanucleus" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/E:/Spark/spark/jars/datanucleus-core-3.2.10.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/E:/Spark/spark/bin/../jars/datanucleus-core-3.2.10.jar."

17/12/06 00:38:33 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/E:/Spark/spark/jars/datanucleus-api-jdo-3.2.6.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/E:/Spark/spark/bin/../jars/datanucleus-api-jdo-3.2.6.jar."

17/12/06 00:39:03 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0

17/12/06 00:39:04 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException

17/12/06 00:39:08 WARN ObjectStore: Failed to get database global\_temp, returning NoSuchObjectException

Spark context Web UI available at http://192.168.194.1:4040

Spark context available as 'sc' (master = local[\*], app id = local-1512500907834).

Spark session available as 'spark'.

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\ version 2.2.0

/\_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0\_152)

Type in expressions to have them evaluated.

Type :help for more information.

scala> val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

<console>:1: error: invalid escape character

val ManUtdPlayers = sc.textFile("E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv")

^

scala> val ManUtdPlayers = sc.textFile("E:\\Chicago\_E\_Drive\\Deepak DOCs\\Study\\Data Analytics\\Sample\_Data\\ManUtdPlayers.csv")

ManUtdPlayers: org.apache.spark.rdd.RDD[String] = E:\Chicago\_E\_Drive\Deepak DOCs\Study\Data Analytics\Sample\_Data\ManUtdPlayers.csv MapPartitionsRDD[1] at textFile at <console>:24

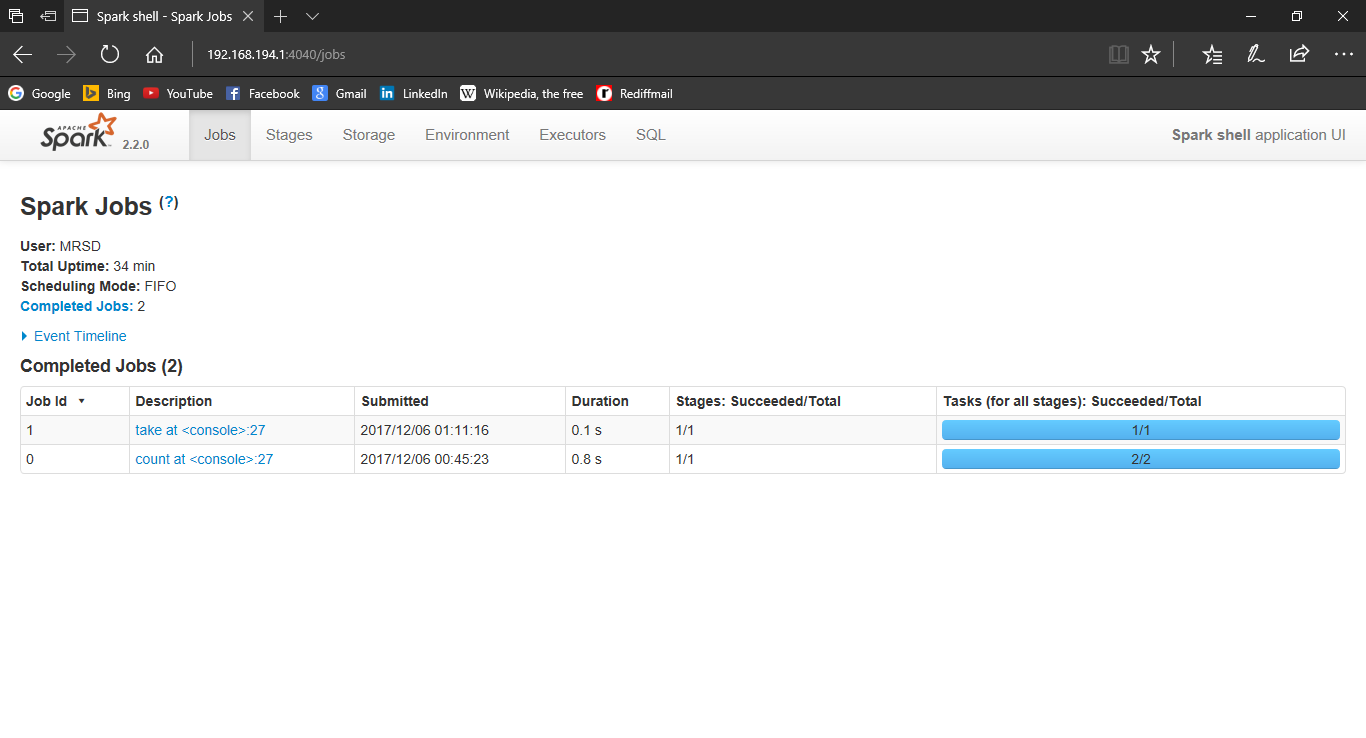
scala> ManUtdPlayers.count()

res0: Long = 547

scala> ManUtdPlayers.take(10)

res1: Array[String] = Array(Peter Abbott,England,Forward,1/10/1953, Stan Ackerley,Australia,Defender,12/7/1942, Ted Adams,England,Goalkeeper,30/11/1906, Alphonso Ainsworth,England,Forward,31/07/1913, Nicky Ajose,England,Forward,7/10/1991, George Albinson,England,Midfielder,14/02/1897, Arthur Albiston,Scotland,Defender,14/07/1957, Reg Allen,England,Goalkeeper,3/5/1919, Arthur Allman,England,Defender,24/12/1890, Alfred Ambler,England,Midfielder,01/07/1879)

# Spark UI running on <http://192.168.194.1:4040>



## Environment

#### Runtime Information

| **Name** | **Value** |
| --- | --- |
| Java Home | E:\Spark\jdk\jre |
| Java Version | 1.8.0\_152 (Oracle Corporation) |
| Scala Version | version 2.11.8 |

#### Spark Properties

| **Name** | **Value** |
| --- | --- |
| spark.app.id | local-1512500907834 |
| spark.app.name | Spark shell |
| spark.driver.host | 192.168.194.1 |
| spark.driver.port | 57440 |
| spark.executor.id | driver |
| spark.home | E:\Spark\spark\bin\.. |
| spark.jars |  |
| spark.master | local[\*] |
| spark.repl.class.outputDir | C:\Users\MRSD\AppData\Local\Temp\spark-d74d283c-2500-4886-8e1e-fb4eb5025e1e\repl-84018871-6492-4bfc-8183-8b3b1baa3f93 |
| spark.repl.class.uri | spark://192.168.194.1:57440/classes |
| spark.scheduler.mode | FIFO |
| spark.sql.catalogImplementation | hive |
| spark.submit.deployMode | client |

#### System Properties

| **Name** | **Value** |
| --- | --- |
| SPARK\_SUBMIT | true |
| awt.toolkit | sun.awt.windows.WToolkit |
| file.encoding | Cp1252 |
| file.encoding.pkg | sun.io |
| file.separator | \ |
| java.awt.graphicsenv | sun.awt.Win32GraphicsEnvironment |
| java.awt.printerjob | sun.awt.windows.WPrinterJob |
| java.class.version | 52.0 |
| java.endorsed.dirs | E:\Spark\jdk\jre\lib\endorsed |
| java.ext.dirs | E:\Spark\jdk\jre\lib\ext;C:\WINDOWS\Sun\Java\lib\ext |
| java.home | E:\Spark\jdk\jre |
| java.io.tmpdir | C:\Users\MRSD\AppData\Local\Temp\ |
| java.library.path | E:\Spark\jdk\bin;C:\WINDOWS\Sun\Java\bin;C:\WINDOWS\system32;C:\WINDOWS;C:\ProgramData\Oracle\Java\javapath;C:\Program Files (x86)\PC Connectivity Solution\;E:\Chicago\_E\_Drive\Deepak DOCs\Study\Rexx\_Clist\_Panel\_Skeleton\Regina\_Latest\Regina;C:\Program Files (x86)\Skype\Phone\;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerShell\v1.0\;C:\Program Files (x86)\ATI Technologies\ATI.ACE\Core-Static;E:\Chicago\_E\_Drive\Deepak DOCs\Study\Python\Python\_installation\Scripts\;E:\Chicago\_E\_Drive\Deepak DOCs\Study\Python\Python\_installation\;C:\Users\MRSD\AppData\Local\Microsoft\WindowsApps;E:\Spark\jdk\bin;E:\Spark\spark\bin;;. |
| java.runtime.name | Java(TM) SE Runtime Environment |
| java.runtime.version | 1.8.0\_152-b16 |
| java.specification.name | Java Platform API Specification |
| java.specification.vendor | Oracle Corporation |
| java.specification.version | 1.8 |
| java.vendor | Oracle Corporation |
| java.vendor.url | http://java.oracle.com/ |
| java.vendor.url.bug | http://bugreport.sun.com/bugreport/ |
| java.version | 1.8.0\_152 |
| java.vm.info | mixed mode |
| java.vm.name | Java HotSpot(TM) 64-Bit Server VM |
| java.vm.specification.name | Java Virtual Machine Specification |
| java.vm.specification.vendor | Oracle Corporation |
| java.vm.specification.version | 1.8 |
| java.vm.vendor | Oracle Corporation |
| java.vm.version | 25.152-b16 |
| line.separator |  |
| os.arch | amd64 |
| os.name | Windows 10 |
| os.version | 10.0 |
| path.separator | ; |
| scala.usejavacp | true |
| sun.arch.data.model | 64 |
| sun.boot.class.path | E:\Spark\jdk\jre\lib\resources.jar;E:\Spark\jdk\jre\lib\rt.jar;E:\Spark\jdk\jre\lib\sunrsasign.jar;E:\Spark\jdk\jre\lib\jsse.jar;E:\Spark\jdk\jre\lib\jce.jar;E:\Spark\jdk\jre\lib\charsets.jar;E:\Spark\jdk\jre\lib\jfr.jar;E:\Spark\jdk\jre\classes |
| sun.boot.library.path | E:\Spark\jdk\jre\bin |
| sun.cpu.endian | little |
| sun.cpu.isalist | amd64 |
| sun.desktop | windows |
| sun.io.unicode.encoding | UnicodeLittle |
| sun.java.command | org.apache.spark.deploy.SparkSubmit --class org.apache.spark.repl.Main --name Spark shell spark-shell |
| sun.java.launcher | SUN\_STANDARD |
| sun.jnu.encoding | Cp1252 |
| sun.management.compiler | HotSpot 64-Bit Tiered Compilers |
| sun.nio.ch.bugLevel |  |
| sun.os.patch.level |  |
| sun.stderr.encoding | cp437 |
| sun.stdout.encoding | cp437 |
| user.country | US |
| user.dir | E:\Spark\spark |
| user.home | C:\Users\MRSD |
| user.language | en |
| user.name | MRSD |
| user.script |  |
| user.timezone | Asia/Calcutta |
| user.variant |  |

# Exit spark-shell

scala> exit

<console>:24: error: not found: value exit

exit

^

scala> exit()

<console>:24: error: not found: value exit

exit()

^

scala> quit()

<console>:24: error: not found: value quit

quit()

^

scala> quit

<console>:24: error: not found: value quit

quit

^

scala> :q

E:\Spark\spark>

# Start-up pyspark

Windows PowerShell

Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\MRSD> pyspark

Python 3.5.1 (v3.5.1:37a07cee5969, Dec 6 2015, 01:54:25) [MSC v.1900 64 bit (AMD64)] on win32

Type "help", "copyright", "credits" or "license" for more information.

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

17/12/30 14:47:44 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/12/30 14:48:26 WARN ObjectStore: Failed to get database global\_temp, returning NoSuchObjectException

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_ / .\_\_/\\_,\_/\_/ /\_/\\_\ version 2.2.0

/\_/

Using Python version 3.5.1 (v3.5.1:37a07cee5969, Dec 6 2015 01:54:25)

SparkSession available as 'spark'.

>>> ManUtdPlayers = sc.textFile("E:\\Chicago\_E\_Drive\\Deepak DOCs\\Study\\Data Analytics\\Sample\_Data\\ManUtdPlayers.csv")

>>> counts = ManUtdPlayers.flatMap(lambda line: line.split(" ")) \

... .map(lambda word: (word, 1)) \

... .reduceByKey(lambda a, b: a + b)

>>> counts.saveAsTextFile("E:\\Chicago\_E\_Drive\\Deepak DOCs\\Study\\Data Analytics\\Sample\_Data\\ManUtdPlayers\_pyspark\_wordcount")

[Stage 0:> (0 + 2) / 2]E:\Spark\spark\python\lib\pyspark.zip\pyspark\shuffle.py:58: UserWarning: Please install psutil to have better support with spilling

E:\Spark\spark\python\lib\pyspark.zip\pyspark\shuffle.py:58: UserWarning: Please install psutil to have better support with spilling

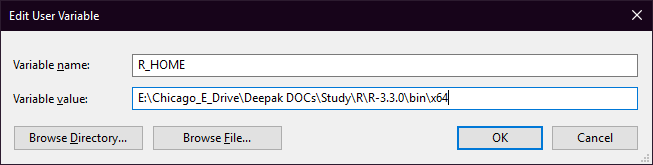
[Stage 1:> (0 + 2) / 2]E:\Spark\spark\python\lib\pyspark.zip\pyspark\shuffle.py:58: UserWarning: Please install psutil to have better support with spilling

E:\Spark\spark\python\lib\pyspark.zip\pyspark\shuffle.py:58: UserWarning: Please install psutil to have better support with spilling

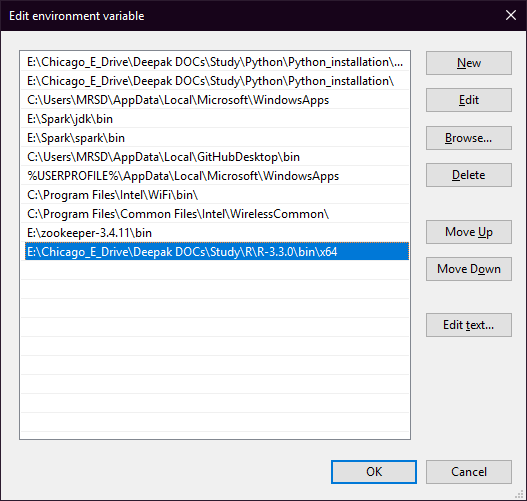
>>> exit()

# Start-up sparkR

Install R and RStudio.



E:\Chicago\_E\_Drive\Deepak DOCs\Study\R\R-3.3.0\bin\x64



Open RStudio.

> Sys.setenv(SPARK\_HOME = "E:\Spark\spark").libPaths(c(file.path(Sys.setenv("SPARK\_HOME"),"R","lib"),.libPaths()))library(SparkR)

Error: '\S' is an unrecognized escape in character string starting ""E:\S"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths(c(file.path(Sys.setenv("SPARK\_HOME"),"R","lib"),.libPaths()))library(SparkR)

Error: unexpected symbol in "Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))library(SparkR)

Error: unexpected symbol in "Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths())).library(SparkR)

Error: unexpected symbol in "Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))

Error: unexpected symbol in "Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark")

> libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))

Error: could not find function "libPaths"

> Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))

Error: unexpected symbol in "Sys.setenv(SPARK\_HOME = "E:/Spark/spark").libPaths"

> .libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))

> library(SparkR)

Attaching package: ‘SparkR’

The following objects are masked from ‘package:stats’:

cov, filter, lag, na.omit, predict, sd, var, window

The following objects are masked from ‘package:base’:

as.data.frame, colnames, colnames<-, drop, endsWith, intersect, rank,

rbind, sample, startsWith, subset, summary, transform, union

> sc <- sparkR.init(master = "local")

Launching java with spark-submit command E:/Spark/spark/bin/spark-submit2.cmd sparkr-shell C:\Users\MRSD\AppData\Local\Temp\RtmpS2HyT5\backend\_port26b41a35745f

Warning message:

'sparkR.init' is deprecated.

Use 'sparkR.session' instead.

See help("Deprecated")

> sqlContext <- sparkRSQL.init(sc)

Warning message:

'sparkRSQL.init' is deprecated.

Use 'sparkR.session' instead.

See help("Deprecated")

> DF <- createDataFrame(sqlContext, faithful)

Warning message:

'createDataFrame(sqlContext...)' is deprecated.

Use 'createDataFrame(data, schema = NULL)' instead.

See help("Deprecated")

>

> head(DF)

eruptions waiting

1 3.600 79

2 1.800 54

3 3.333 74

4 2.283 62

5 4.533 85

6 2.883 55

> localDF <- data.frame(name=c("John", "Smith", "Sarah"), age=c(19, 23, 18))

> df <- createDataFrame(sqlContext, localDF)

Warning message:

'createDataFrame(sqlContext...)' is deprecated.

Use 'createDataFrame(data, schema = NULL)' instead.

See help("Deprecated")

> printSchema(df)

root

|-- name: string (nullable = true)

|-- age: double (nullable = true)

> path <- file.path(Sys.getenv("SPARK\_HOME"), "examples/src/main/resources/people.json")

>

> peopleDF <- jsonFile(sqlContext, path)

Warning messages:

1: 'jsonFile(sqlContext...)' is deprecated.

Use 'jsonFile(path)' instead.

See help("Deprecated")

2: 'f' is deprecated.

Use 'read.json' instead.

See help("Deprecated")

> printSchema(peopleDF)

root

|-- age: long (nullable = true)

|-- name: string (nullable = true)

> registerTempTable(peopleDF, "people")

Warning message:

'registerTempTable' is deprecated.

Use 'createOrReplaceTempView' instead.

See help("Deprecated")

> teenagers <- sql(sqlContext, "SELECT name FROM people WHERE age >= 13 AND age <= 19")

Warning message:

'sql(sqlContext...)' is deprecated.

Use 'sql(sqlQuery)' instead.

See help("Deprecated")

> teenagersLocalDF <- collect(teenagers)

> print(teenagersLocalDF)

name

1 Justin

> allPeople <- sql(sqlContext, "SELECT \* FROM people")

Warning message:

'sql(sqlContext...)' is deprecated.

Use 'sql(sqlQuery)' instead.

See help("Deprecated")

> printAllPeople <- collect(allPeople)

> print(printAllPeople)

age name

1 NA Michael

2 30 Andy

3 19 Justin

> sparkR.stop()